Why Does It Matter How NAEP Results Are Produced?

You may be thinking, 'Why do I need to understand this?' Isn't it enough to know what the results are?

In order to understand and to explain to others what the results mean, one needs to know what kind of information the National Assessment of Educational Progress data provide.

You may be already aware that not every 4th or 8th grade student in Maine takes part in the assessment and that no student sees the entire assessment for a subject at either grade level.

Also, although 12th graders in Maine do participate in the assessments, the state receives results only for 4th and 8th grade. NAEP was originally designed to be a measure of the nation's progress as a whole towards ambitious educational goals.

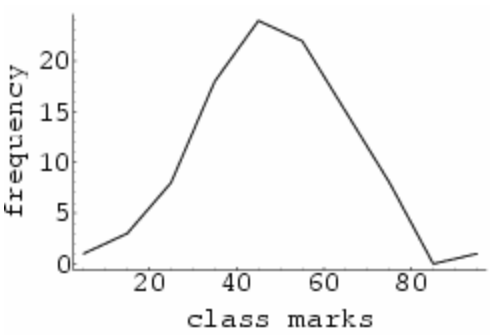
State levels NAEP results have been available only for a little more than a decade, and administration or Reading and Mathematics at 4th and 8th grade in every state was required by law only with the implementation of the No Child Left Behind Act several years ago. However, Maine has participated in NAEP almost since it began.

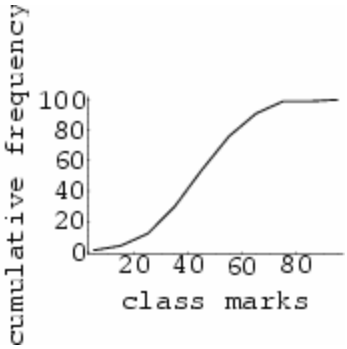
It has not been financially practical to administer the assessments to every student in the nation at a grade level; instead, the NAEP design calls for a sampling of students and of assessment items. Participating students see only parts of a very large test designed to cover the entire subject being assessed. This is necessary to make the NAEP administration as little of an intrusion upon school schedules as possible while providing an adequate range of items to provide a fair assessment of student ability across the nation.

Out of the pieces of raw data gathered from this piecemeal administration of the assessment, NAEP results are generated by a statistical modeling procedure known as Item Response Theory (IRT).

Note: Thanks to New Hampshire NAEP State Coordinator David Gebhardt for assistance with some of the graphics below.

Bell Curves & Ogives: Grading vs. Assessing





An IRT plot is an Ogive, a continuous cumulative frequency curve, such as the one illustrated above in the right figure (otherwise known as an S-curve).

The word is of uncertain origin; some dictionaries trace it to the French 'auge' meaning 'trough' or the Latin 'augere' meaning 'to increase.' Others link it to an ancient Arabic astrological word for the 'highest point.'

The graph on the left above shows a curve resulting from the plot of grades versus the number of times they were given in a class. Notice that very high and very low scores are less common than scores given in the mid-range, resulting in a sort of bell-shaped curve.

The S shaped curve on the right is the result of plotting the same scores against an accumulation of scores, so that any point plotted along the line would represent all the scores given up to the number indicated; grades of 60 and below appear to account for a little more than 80% of the scores given in the class.

On the left we are looking at the frequency with which different scores are given by a teacher; on the right we are looking at a statistical interpretation of the teacher's behavior in giving grades. On the left, we have some raw information; on the right we have the beginning of an explanation of what this information means.

From this point, most discussions of IRT quickly become very abstract. Perhaps if we start with the concrete, we can keep our feet on the ground for this exploratory discussion.

The most important point for our discussion so far is that the raw data collected in NAEP assessments is nothing like the scores individual students receive on MEA assessments. No students receive individual scores on NAEP assessments; in fact scores cannot even be generated for schools or districts because of the wide distribution of different items to a relatively small number of students in the state.

What Do Maine NAEP Scores Mean?

The results we report for Maine NAEP are taken from the NAEP Data Tool, which is available to the general public through the Maine Department of Education Web Site.

Accessing the NAEP Data Tool
from http://www.state.me.us/education/homepage.htm

- Click on **Standards & Assessment** in the upper left corner of the page.
- Click on **National Assessment of Educational Progress** in the pull-down window
- Scroll down to the bottom of the page, and click on **NAEP Research e-Center.**
- Click on **NAEP Data on the Web** on the left halfway down the page.

Following this path to the NAEP Data Tool will take you through some very interesting materials related to assessment. When you get to your destination, we recommend that you bookmark it for easy access in the future.

The NAEP Data Tool reports group scores and percentages of students at the different achievement levels for individual states and for the nation as a whole.

A score report based upon gender groups in Maine looks like this:

Maine/Mathematics Composite/Grade 4/2003 and 2000
Gender of student as taken from school records *[GENDER]*
Average Scale Score and Row Percentage (with Standard Errors in Parentheses)

OVERALL

		Male		Female	
		Average Scale	Row	Average Scale	Row
	Year	N Score	Percentage	Score	Percentage
Total	2003	2879239(0.9)	51%(0.8)	236(0.9)	49%(0.8)
	2000	2202232(1.2)	51%(1.0)	227(1.3)	49%(1.0)

NOTE: The NAEP Mathematics scale ranges from 0 to 500. Observed differences are not necessarily statistically significant.

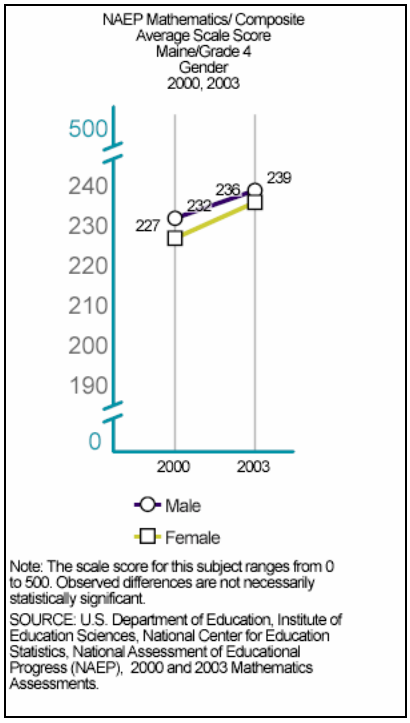
SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 and 2000 Mathematics Assessments.

In the table above, "N" represents the number of students sampled in Maine for 4th grade Mathematics results. The "Average Scale Score" is a number generated by IRT. The "Row Percentage" is the proportion of the category identified (in this case, gender) in the overall population (N) sampled. The numbers in parentheses are sampling errors

associated with all surveys; here 239 given as the score for males in 2003 actually represents a range of possible scores of 238.1 to 239.9, but we will see below that interpretation of the range for a given IRT score can be even larger.

Notice that between 2000 and 2003 the number of students assessed (N) increased and the standard error (in parentheses after the scores) decreased. Generally, the large the sample size for an assessment, the smaller the margin of error.

In presenting this table, the Data Tool is not generating real-time statistical analyses; it is displaying tables previously generated for different subgroups by NAEP statisticians. It will also generate charts:



These tables and charts can be cut and pasted to your own documents, or they can be printed directly. The Data Tool will also generate maps for interstate comparisons of population subgroups.

> *The NAEP State Coordinator for Maine is available for school or district presentations on uses of the NAEP Data Tool and the NAEP Question Tool, which is a database of item classification, content, and scoring rubric information. Contact John Kennedy at 207-624-6636 or john.kennedy@maine.gov for more information.*

Again, we don't actually get to look at raw NAEP data in these tools; if we did, it would not make much sense at first anyway.

Raw NAEP Data

A disk of raw data obtained from the National Center of Education Statistics contains the separate pieces of information used to generate NAEP results, including ***student-level records*** and ***data codebooks***. Here one quickly finds oneself in another kind of environment than the one of the NAEP Data Tool.

Among other data, a NAEP ***student-level record*** contains *plausible values*:

Contents of Student-Level Record

- Identification information & sample indicators
- Population and sample-based ***weights***
- Reporting categories & derived variables
- *Plausible values* (ability estimates), also known as ***thetas***
- Questionnaire responses

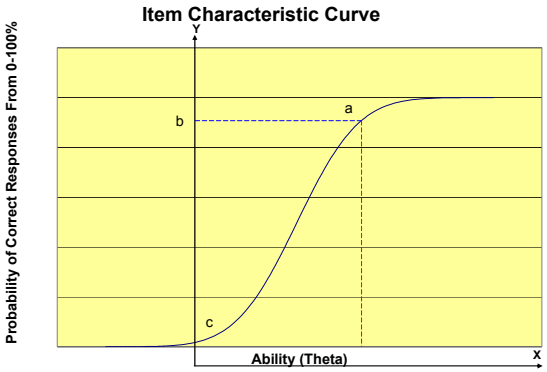
Among other data, a ***data codebook file*** contains information about *IRT parameters*:

Contents of Codebook File

- Record Layout (descriptions of formats of data tables)
- Data Codebooks (item location in test books, data values and their frequencies, **IRT parameters**, scoring keys)

IRT parameters and **thetas** are the foundation of NAEP average scaled scores for states.

In Item Response Theory (IRT), the probability that a student with a designated ability level (called a ***theta***) will correctly answer an item having three specified parameters associated with S-shaped probability curves (Ogives) is calculated using complex statistical formulae.



These **item parameters** are designated as *a, b, and c* in IRT. Notice the location of the label "theta" in the diagram above. The other axis is scaled in probabilities of 0% to 100%.

Thetas are not associated with ethnicity, gender, region, economic status or any other population indicator; thetas are abstract indicators of ability regardless of anything other than the difficulty of the items.

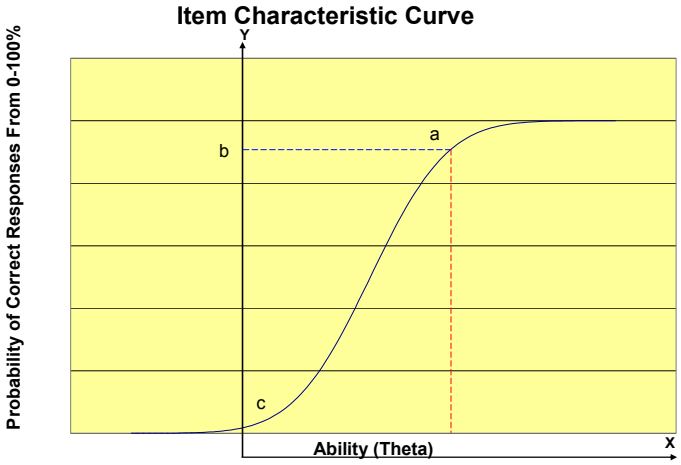
Weights (refer to *Contents of Student-Level Record* above) not thetas are used to provide subgroup scores (say, for poor males in rural areas of a certain state).

Thetas do not change over time; they are said to be *invariant,* a characteristic that permits the evaluation of individual items before they are actually administered to a test population.

IRT is a process of statistical modeling of test-taker behavior. This model allows the estimation of sampling errors based upon actual test item performance rather than the classical method of assuming that error averages out across a population of test-takers.

This last point has implications for Computer Adaptive Testing because it allows test makers to select items that most effectively measure the ability of test-takers at specific ability levels.

IRT Parameters



The b parameter designates the difficulty of the item in itself and is associated with the probability of the number of students of differing abilities being able to answer it correctly. It is the point on the Y-axis which is on a horizontal line with a specific point on the curve.

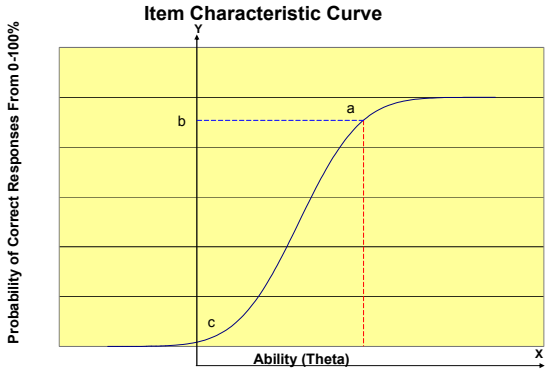
As we move out the X-axis, we include more *thetas* or more levels of ability. So an item associated with success at a lower theta level is also associated with success at all the higher theta levels. This, again, is the cumulative curve of the type represented in the graph on the right at the top of this document. It is an interpretation of the performance of an item in a test. The probabilities of success on this item are indicated on the Y-axis.

The *a* parameter gives a numerical value to the ability of the item to discriminate between students of higher and lower ability. It is the slope (angle of ascent) of the curve at the point of the *a* parameter. Notice that S-curves flatten out at the top and bottom of these graphs. This is because a test-taker associated with a very low theta is not likely to be successful on most items and a test-taker associated with a very high theta is likely to be successful on nearly any item.

The *c* parameter is the point at which the S-curve intercepts the Y-axis (where there is no theta). This is sometimes called the *guessing parameter* since even a student of no ability theoretically would have some probability of answering correctly on the item — presumably by guessing.

The item parameters are calculated from item's actual performance and combined with the examinee ability estimates (*thetas*) to create a model of the probability of students of differing abilities providing a correct answer to the item.

This results in the *Item Characteristic Curve* (ICC) for an item. ICC's are the result of plugging different possible values into an equation containing *a, b, c,* and *theta* until a solution to the equation meeting the requirements of probability theory emerges; i.e., until a S-curve results.



One can experiment with different IRT parameters and their resulting S-curves on the Internet at http://edres.org/scripts/cat/genicc.asp.

ICC's are combined to create Test Characteristics Curves (TCC's), which predict how students who did not take the test but have similar theoretical ability levels (*thetas)* would have scored on the test had they taken it. One notices that S-curves for TCC's also flatten out for items that are very easy or very difficult; this means that very easy items give us

very little statistical information about more able students and very difficult items give us very little statistical information about less able students.

While it relatively easy to see where the item parameters are coming from if one looks at an actual S-curve plot on a graph, the reality of ***theta*** is a bit more difficult to grasp.

Is a theta something like an IQ score or a performance rating by a supervisor? Not exactly; it is a variable in an equation that creates a model of test-takers interacting with items. It is part of a simulation of a situation that leads to a conclusion about the data an item provides. And it allows us to make predictions about the performance of students assessed by NAEP on those items in the assessment they never see, as well as predicting the performance of students not assessed but with characteristics similar to those students that were.

IRT score generation, then, is a juggling of several values related to each item that results in the *maximum likelihood* that a specific ability level is associated with a probability of providing the correct response to an item.

These item-level statistics are combined for all the items in an assessment and weighted to provide a profile of a specific student subgroup achieving a specific score range.

IRT does not generate scores in the same way that a classroom test would because the IRT score is not the percentage of right answers for a specific student. NAEP scores cannot be reliably generated for individual students; they are interpretations of trends in large populations of students.

This is why all NAEP results are reported with the caution that "observed [score] differences are not necessarily statistically significant." Because IRT scores for NAEP are generated by a statistical model, sets of them must be tested statistically to see if they differ among themselves.

This is more than the issue of margin of error; it is a feature of the statistical universe in which they exist. The NAEP Data Tool will perform a real-time *test of statistical significance* upon a table of data it generates. This test can be accessed through the User Options pull down menu at the top of the screen.

See ***Directions for Accessing the Data Tool*** *above.*

Demonstrations of the range of features available to the public on the NAEP Data Tool can be arranged with the NAEP State Coordinator for Maine. NAEP will soon be releasing a new version of the Data Tool, allowing for more complex investigations of the factors in and out of school that influence student performance.

The NAEP 2005 administration of Reading, Mathematics, and Science assessments is scheduled for January 24 to February 18 in Maine. More schools than in the past will be selected for participation, which is required under the *No Child Left Behind* Act. School selections by NAEP will be announced in August.